عنوان :

رگرسیون لاسوی لجستیک بیزی با انتخاب توزیع پیشین برای پارامتر تنظیم در داده‌های بعد بالا

توسط :

ابوالفضل حسین نتاج

اساتید راهنما :

دکتر عباس بهرامپور | دکتر محمدرضا بانشی

شماره پایان نامه: ۱۰/۸/۱/۲ سال تحصیلی : ۱۳۹۹-۱۳۹۸

بسمه تعالی

تاریخ .....:

شماره.....:

پیوست.....:

صورتجلسه دفاع از پایان نامه

دانشگاه علوم پزشکی کرمان

معاونت تحصیلات تکمیلی دانشگاه

جلسه دفاعیه پایان نامه تحصیلی آقای ابوالفضل حسین نتاج دانشجوی دکتری تخصصی (Ph.D) رشته آمار زیستی تحت عنوان" رگرسیون لاسوی لجستیک بیزی با انتخاب توزیع پیشین برای پارامتر تنظیم در داده های بعد بالا " با حضور هیئت داوران برگزار گردید و پایان نامه به شرح ذیل مورد تأیید قرار گرفت.

در ساعت .....۱۲..... روز .....چهارشنبه..... مورخ .....۹۸/۸/۱..... با حضور اعضای محترم هیات داوران به شرح ذیل:

سمت	نام و نام خانوادگی	امضا
الف:استاد(ان) راهنما	۱ آقای دکتر عباس بهرامپور ۲ آقای دکتر محمدرضا باتشی	
ب: استاد(ان) مشاور		
ج: عضو هیات داوران (داخلی)	آقای دکتر یونس جهانی	
ج: عضو هیات داوران (داخلی)	خانم دکتر مقدمه میرزایی	
د :عضو هیات داوران (خارجی)	آقای دکتر حسین محجوب	
د :عضو هیات داوران (خارجی)	آقای دکتر آوات فیضی	
ه :نماینده تحصیلات تکمیلی	آقای دکتر محمدرضا باتشی	

تشکیل گردید و ضمن ارزیابی به شرح پیوست با درجه .....عالی..... و نمره .....نوزده و نیم (۱۹/۵)..... مورد تأیید قرار گرفت.

معاون آموزشی

چکیده

مقدمه و اهداف: دو معضل داده ها برای تحلیل آماری کم بودن حجم نمونه و وجود همخطی مـی باشـد. در سال های اخیر رگرسیون لاسو به عنوان یکی از راهکارهای این مشکلات مورد توجه قـرار گرفتـه اسـت. یـک عامل مهم در رگرسیون لاسو، پارامتر تنظیم می باشد که نقش مهمی در نتایج ایـن رگرسـیون دارد. راه هـای متنوعی برای برآورد این پارامتر پیشـنهاد شـده اسـت کـه در ایـن مطالعـه بـه دو نـوع آن (بیـزی و غیربیـزی) پرداخته شد. در نهایت عملکرد چهار مدل لاسو لجستیک مورد مقایسه قرار گرفـت. بعضـی از ایـن مـدل هـا برای اولین بار در رگرسیون لجستیک بکار گرفته شد.

روش ها: این مطالعه از نوع مدل سازی می باشد. در این پایان نامه به مقایسه عملکرد چهار مدل بـا سـاختار متفاوت توزیع پسین پرداخته شد. مدل ها شامل: مدل لاسو لجستیک معمولی که براسـاس منطـق بیـزی و با انتخاب توزیع پیشین لاپلاس برای ضرایب رگرسیونی و مقدار ثابت برای پارامتر تنظیم مورد اسـتفاده قـرار گرفت(مدل ۱)، مدل بیزی با توزیع پیشین گاما برای مربع پارامتر تنظـیم و توزیـع هـای نرمـال و گامـا بـرای سایر پارامترها (مدل ۲)، مدل بیزی با توزیع پیشین گاما برای پارامتر تنظیم و توزیع هـای یکنواخـت و گامـا بـرای سـایر پارامترهـا (مـدل ۳) و مـدل بیـزی بـا توزیـع هـای پیشـین گامـا بـرای پـارامتر تنظـیم و ضـرایب رگرسیونی(مدل ۴) بودند. پارامتر تنظیم در مدل ۱ از طریق روش اعتبارسـنجی متقابـل بدسـت آمـد. بـرای سایر مدل ها (بیزی) توزیع های پیشین مختلف برای پارامتر تنظیم در نظر گرفته شد.

بمنظور مقایسه عملکرد مدل ها از داده های بیماران سرطان معده (شـامل ۳۳۹ بیمـار مبـتلا بـه سـرطان معده در دو بیمارستان شهر کرمان در طی سال های ۱۳۸۰ تا ۱۳۹۴) و داده های شبیه سـازی شـده (پـنج سناریو با حجم نمونه و ساختار همبستگی های مختلف) استفاده شد. شاخص های عملکـردی مـدل هـای مختلف بر روی داده های قسمت آزمایشی محاسبه شدند. همچنین در داده های سـرطان معـده عـلاوه بـر مقایسه عملکرد مدل ها، عوامل خطر در مرگ بیماران شناسایی شدند.

یافتـه هـا: در داده هـای سـرطان معـده، ٦٣/٧ % مـرد بودنـد. میـانگین و انحـراف معیـار سـن بیمـاران ١٤/٥٢±٦٢/٨٤ سال بود. در پایان مطالعه، ١٩٥ نفر از افراد مبتلا به سرطان معده فوت کـرده‌انـد. در ایـن داده ها، مدل ٢ و سپس مدل ١ دارای بهترین عملکرد بودند. میانه پارامتر تنظیم در مدل ٢ دارای بزرگترین مقدار بود. بصورت کلی متغیرهای جنسیت، شکل تومور و مدت زمان تشخیص بیماری دارای ارتباط معنـی دار با مرگ بیماران شناخته شدند.

در داد های شبیه سازی شده با افزایش حجم نمونه، عملکرد مدل ها بر اساس شـاخص هـای Accuracy و Precision بهبود یافت. همچنین نتیجه گرفته شد هـر چـه همبسـتگی بـین متغیرهـا افـزایش مـی یابـد، عملکرد مدل ها ضعیف می شوند. بصورت کلـی در سـناریو هـای مختلـف مـدل هـای ١ و ٢ دارای بهتـرین عملکرد بودند. همچنین در همه سناریوها شاخص Precision مدل ٢ بهتر از سایر مدل ها بود.

بحث و نتیجه گیری: ترتیب عملکرد مدل ها در این مطالعه عبارتند از: مدل ٢، مدل ١، مـدل ۴ و در نهایـت مدل ٣.

با این رابطه می توان نتیجه گرفت انتخاب مناسب ساختار برای پارامترهای مدل بسیار مهم می باشـد. در بعضی از سناریوهای شبیه سازی شده، مدل ٢ دارای بهترین عملکرد بود و در بعضی از مـوارد مـدل ١ بهتـر از سایر مدل ها بود. با توجه به پیچیدگی کمتر و سرعت اجرای بیشتر، می توان نتیجه گرفت که مدل لاسـو معمولی (مدل ١) مورد استفاده در این مطالعه اختلاف قابل توجهی با مدل بیزی ٢ ندارد.

کلمات کلیدی: رگرسیون لاسو، تحلیل بیزی، پارامتر تنظیم، سرطان معده

Abstract

Background & Objective: Two main issues that challenge the practice of statistical analysis are small sample size and multi-collinearity. In the recent years, lasso regression is proposed to tackle these issues. In the Lasso regression, tuning parameter has a key role on the estimated regression coefficients, model prediction and goodness of fit. For estimating tuning parameter, various methods have been proposed that two types of them (Bayesian and non-Bayesian) have been discussed in this study. Finally, we compared the performance of four Bayesian and non-Bayesian logistic models using real data and simulation.

Methods: In this modeling study, we compared the performance of four models including one ordinary Lasso logistic model (Model 1) and three Bayesian logistic models (Models 2 to 4) with different posterior distribution structures. In Model 1, the tuning parameter was obtained through the cross validation method. For other models (Bayesian), different prior distributions were considered for the tuning parameter.

In order to compare the performance of the models, gastric cancer data (including 339 gastric cancer patients in two hospitals of Kerman city between 2001 and 2014) and simulated data (five scenarios with different sample size and correlation structure) were used. Performance indices of different models were computed on the test dataset. Also, the influential risk factors on the mortality of the gastric cancer patients were identified.

Results: In gastric cancer data, 63.7% of patients were male. The mean ± SD for age was 62.84 ±14.52 years. The number of death due to gastric cancer at the end of the study were 195 person. In

this data, Model 2 and then Model 1 had the best performance. The median of the tuning parameter in Model 2 was maximum. In totally, the variables gender, tumor shape and date of diagnosis had statistical significant on the mortality.

In the simulated data, the performance of the models improved based on Accuracy and Precision indices is improved as the sample size increased. In addition, the performance of the models decreased when the correlation between the variables increased. Generally in the different scenarios, Models 1 and 2 had the best performance. Also, the precision index of Model 2 in the all scenarios was better in compared to the other models.

Conclusion: The order of performance of the models in this study is as follows:

Bayesian Model 2> Non-Bayesian Model 1> Bayesian model 4> Bayesian model 3

According to the models performance order, the selection of proper structures for model parameters is very important. In the some simulated scenarios, the best performance is related to model 2 and in some other scenarios model 1 had the best performance. Therefore, due to the lower complexity and higher operating speeds, it can be concluded that the non-Bayesian (ordinary) Lasso model used in this study had not notable different from the Bayesian Model 2.

Keywords: Lasso regression, Bayesian analysis, Tuning parameters, Gastric cancer

Kerman University of Medical Sciences

Faculty of Health

In Partial Fulfillment of the requirements for the Degree Ph.D in Biostatistics

Title :

Bayesian Lasso Logistic Regression with prior distribution for tuning parameter in high dimensional data

By

Abolfazl Hosseinnataj

Supervisor :

Dr Abbas Bahrampour | Dr Mohammad Reza Baneshi

Thesis No : 10/8/1/2 Year : 2019